Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Cognitive Account of Advocacy Bias

Moshe Hoffman

Hoffman.Moshe@gmail.com

Independent Researcher https://orcid.org/0000-0001-7874-0074

Evan Russek

Princeton University https://orcid.org/0000-0002-2929-6912

Thomas Griffiths

Princeton University https://orcid.org/0000-0002-5138-7255

Article

Keywords:

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11423613/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

A Cognitive Account of Advocacy Bias

Moshe Hoffman[1,*], Evan Russek[2], Tom Griffiths[3]

[1]Independent Researcher, Boston, MA, USA

[2]Department of Psychology, Hunter College, City University of New York, New York, NY, USA

[3]Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA

[*]Corresponding author: hoffman.moshe@gmail.com

Abstract

Advocates often come to believe their own arguments. We provide evidence for a two-step cognitive account: constructing a case selectively increases processing of useful evidence, and advocates subsequently fail to fully adjust for the resulting bias in processed evidence. Across six preregistered, mostly incentivized experiments using stylized sampling tasks and naturalistic legal and policy tasks ($N = 1{,}361$), selectively presented evidence shifted non-advocates' judgments even when its selection was known; among advocates, random advocacy assignment shifted judgments and produced asymmetric updating—stronger responses to evidence supporting their assigned positions; both effects attenuated after accounting for differential processing. Together, these results show how selective processing and incomplete adjustment can explain advocacy-induced belief change, asymmetric updating, and the influence of selected evidence on informed audiences, without requiring advocacy-aligned beliefs to provide psychological or strategic benefits.

Advocacy bias—the tendency for people's beliefs to shift toward positions they advocate—is well documented and potentially consequential. Advocacy bias has been demonstrated in mock negotiations,[1] randomized debate assignments,[2] incentivized advice-giving contexts,[3] and simulated legal advocacy.[4] Advocacy bias may contribute to divergent political beliefs,[5,6] overconfidence,[7,8] and beliefs that favor one's in-group[9,10] or the prevailing social system.[11,12]

Prominent explanations emphasize the psychological or strategic value of advocacy-aligned beliefs. Such beliefs may be psychologically valuable because they increase the prospect of vindication or align with personal or group interests.[13,14] They may be strategically valuable because believing one's own claims can facilitate persuasion by making advocates more convincing.[8,15–17] Despite their differences, these explanations share a belief-utility logic: advocacy-aligned beliefs are valuable to hold, whether for psychological or strategic reasons.

In contrast, we consider a two-step cognitive account that does not require such belief utility. First, constructing a case changes which evidence becomes cognitively available: advocates

have reason to search for, attend to, elaborate on, retain, and communicate evidence useful for their assigned case. Second, when advocates subsequently evaluate the issue, they do not fully adjust for the resulting bias in the evidence most readily available to them.

This account generates three additional predictions. First, selectively available evidence should bias even external evaluators who do not benefit from the resulting belief. Second, advocates should exhibit a form of asymmetric updating[18–23]—disproportionate responsiveness to evidence useful for constructing their assigned case. Third, advocacy-induced belief change and this asymmetric responsiveness should attenuate after accounting for differences in evidence processing.

This account also combines established features of cognition. Goals and task demands direct attention and memory toward task-relevant information, and information that receives greater attention and elaboration becomes more available and influential in later judgment.[24–31] At the same time, people often fail to adjust fully for selection in the information available to them, perhaps because the appropriate adjustment is uncertain or costly to compute.[32–39]

We test this account across six preregistered, mostly incentivized experiments. Studies 1 and 2 isolate adjustment to known selection processes in abstract sampling tasks. Studies 3 and 4 test whether selectively presented evidence continues to shift informed audiences in legal and policy settings, first when selection is imposed experimentally and then when it is generated by an advocate. Studies 5 and 6 randomly assign participants to advocate opposing sides of legal and policy disputes, independently randomize their evidence exposure, and measure evidence processing and belief change.

Results

Study 1: Incomplete adjustment under rank-based selection

We first tested whether people adequately correct for a rule-based process that selectively determines which evidence they observe.

Participants ($N = 50$) repeatedly estimated the mean height of 10 sticks drawn independently from a uniform distribution over the integers 1–100. On each trial, they saw between one and nine sticks selected according to one of three known rules: the tallest sticks, the shortest sticks, or a random subset (Figure 1a,b). Each participant completed all 27 combinations of selection rule and number revealed and received an accuracy bonus.

Participants clearly recognized that the visible sticks were selected rather than representative (Figure 1c). We defined the **Naive** benchmark as the mean of the visible sticks, ignoring selection. With five sticks hidden, estimates were 21.5 points below this benchmark when the tallest sticks were visible and 18.7 points above it when the shortest were visible; these adjustments increased as more sticks were hidden (both $p < .001$).

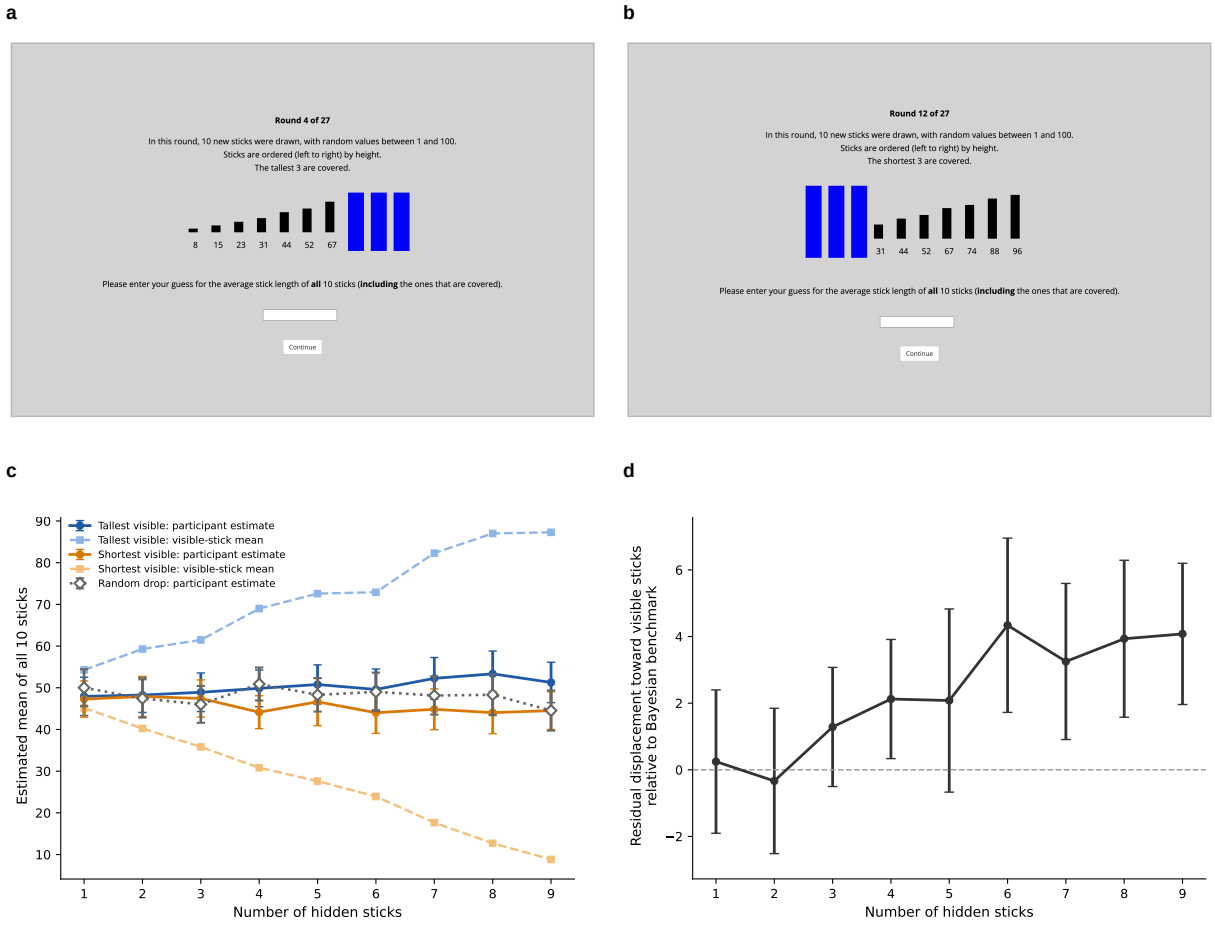


Figure 1: **Incomplete adjustment when the highest or lowest values are hidden. a–b**, Example participant judgment pages with the tallest three or shortest three sticks covered; participants estimated the mean of all 10 sticks from the visible values and covered positions. **c**, Mean participant estimates as a function of the number of hidden sticks. Dashed lines show the corresponding visible-stick means—the Naive benchmark—in the selective-censoring conditions. Participants adjusted away from those means, but estimates remained ordered in the direction of the selectively visible evidence. **d**, Mean residual from the full-correction (Bayesian) benchmark after first averaging the two selective-censoring conditions within participant. Positive values show the remaining pull of the selectively displayed sticks; zero indicates complete correction. In **c–d**, error bars are normal-approximation 95% confidence intervals calculated from participant-level observations. $N = 50$ participants.

Nevertheless, the selection rule continued to affect estimates even though the complete sets of sticks were generated from the same distribution in every condition. With five sticks hidden, estimates were 2.17 points higher when the tallest sticks were visible than under Random Drop (95% CI [0.06, 4.28], $p = .022$) and 2.43 points lower when the shortest were visible (95% CI [−4.23, −0.64], $p = .004$). Thus estimates remained ordered in the direction of the selectively visible evidence.

We next compared estimates with a **Bayesian** benchmark: the expected mean of all 10 sticks conditional on the revealed values, their ordered positions, and the known selection rule (Figure 1d). Because this benchmark incorporates the sampling and selection processes, its expected value is 50.5 under every rule. Formal derivations are reported in the Supplementary Information. Deviations from this benchmark increased in magnitude as more sticks were hidden in both selective-censoring conditions (both $p \leq .033$).

To measure the fraction of the required correction participants made, we also conducted a non-preregistered analysis estimating how strongly their adjustment away from the Naive benchmark tracked the adjustment required to reach the exact Bayesian benchmark. A coefficient $\lambda = 0$ indicates no systematic correction and $\lambda = 1$ complete correction. Pooling the two selective-censoring conditions, participants made approximately 89% of the required adjustment ($\lambda = .891$, 95% CI [.839, .942], $p < .001$ for $\lambda < 1$); adjustment was significantly incomplete in both the Tallest-visible ($\lambda = .872$, $p = .011$) and Shortest-visible conditions ($\lambda = .897$, $p = .035$). Participants therefore responded strongly to the known selection process but remained systematically influenced by which evidence was selected for them to observe.

Study 2: Incomplete adjustment under threshold-based selection

Study 1 examined rank-based selection: a fixed number of the observations most favorable to one side was revealed. Advocacy and many other real-world selection processes more closely resemble threshold-based selection, in which every observation sufficiently favorable to one side enters the record. We therefore tested the same incomplete-adjustment prediction under threshold-based selection. To better approximate the uncertainty characteristic of naturalistic settings, we varied participants' knowledge of the selection process. We also replaced continuous estimation with inference about which of two possible states had generated the evidence, paralleling the competing conclusions in the advocacy studies that follow.

Participants ($N = 116$) completed 30 rounds. On each round, either a Low Machine (mean 230) or High Machine (mean 270; both $SD = 50$) generated a sample. Participants observed a record produced by one of three known rules: **Only High** recorded values above a cutoff, **Only Low** recorded values below it, and **Random Drop** omitted values independently of magnitude. They then estimated the probability that the High Machine had generated the record (Figure 2a,b).

The three blocks introduced increasing uncertainty. In Block 1, participants knew that five values had been sampled and that the cutoff was 250. In Block 2, the number sampled varied and was unobserved. In Block 3, both the number sampled and a cutoff drawn between 230 and 270 were unobserved. Before each block, participants received instructions and could generate unlimited examples of the sampling and recording process.

We again defined a **Naive** benchmark that treated the recorded values as an uncensored sample and a **Bayesian** benchmark that incorporated how the record was produced, integrating over the unobserved sample size and, in Block 3, the cutoff. Block-specific calculations are given in the Supplementary Information.

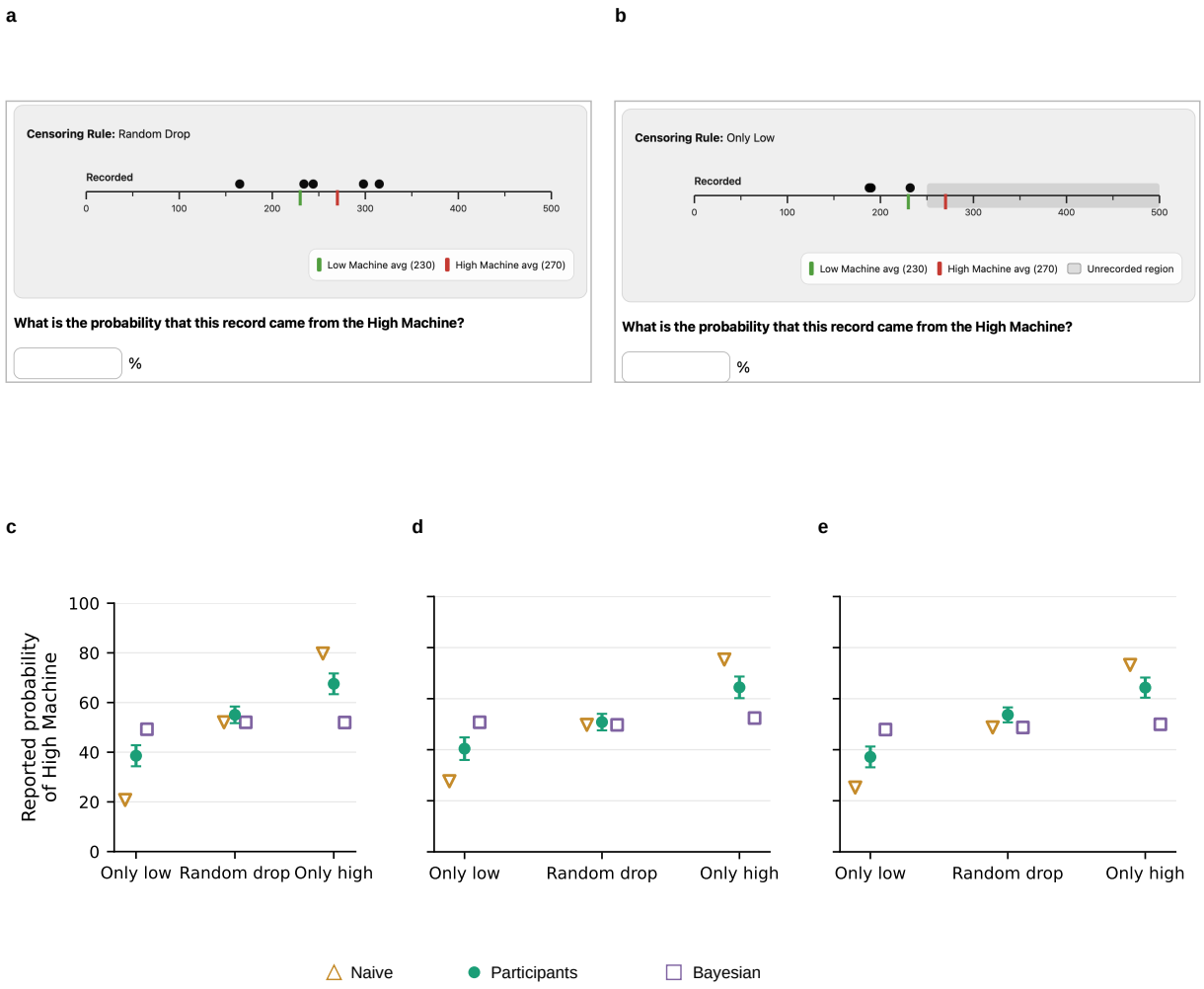


Figure 2: **Incomplete adjustment when only values above or below a cutoff are shown. a–b**, Example participant judgment pages under Random Drop and Only Low from Block 1; the gray region in **b** marks values above the cutoff, which the Only Low rule excluded from the record. Participants saw the recorded values and censoring rule and estimated the probability that the record came from the High Machine. **c–e**, Mean reported probability compared with the Naive and Bayesian benchmarks in Blocks 1–3. The Naive benchmark ignores the censoring rule; the Bayesian benchmark also accounts for how the record was censored. Participant points are means of participant-level condition means; their error bars are normal-approximation 95% confidence intervals across participants. Benchmark symbols show means for the same realized records. Across blocks, judgments in selectively censored conditions lie between the benchmarks, indicating substantial but incomplete correction. $N = 116$ participants.

Across every block and both selective-censoring rules, mean judgments fell between the Naive and Bayesian predictions (Figure 2c–e). All six preregistered proportional-adjustment estimates were significantly above zero and below one (range .286–.428; all $p < .001$). Pooling

across blocks and selective rules, participants made approximately 37% of the required correction ($\lambda = .369$, 95% CI [.282, .456]; $p < .001$ relative to both zero and one).

Because the Naive and Bayesian benchmarks coincide under Random Drop, Study 2 also allowed a complementary test. Absolute deviations from the Bayesian benchmark were greater under Only High (+.065, 95% CI [.045, .085]) and Only Low (+.076, 95% CI [.055, .098]) than under Random Drop (both $p < .001$). Together, Studies 1 and 2 show incomplete adjustment under both rank- and threshold-based selection, including when the sampling process was fully specified.

Study 3: One-sided legal evidence influences audience judgments despite disclosure

We next moved from stylized selection tasks to a naturalistic legal setting, testing whether one-sided evidence would influence judgments even when participants were explicitly informed of its selective presentation.

Participants ($N = 399$) evaluated one of three legal cases, reported initial beliefs about the factual claims at issue and the overall verdict, reviewed evidence, and reported the same judgments again (Figure 3a). They were assigned to see either one fact from each side per claim (the Both-Sides condition) or only plaintiff- or defendant-favoring facts. In the one-sided conditions, participants were either explicitly told that opposing evidence was missing and reminded to account for it, or were not given this disclosure.

Even with disclosure, one-sided evidence shifted both claim and verdict judgments toward the side shown (Figure 3b,c). On the 0–100 judgment scales, relative to the Both-Sides condition, disclosed plaintiff-only evidence shifted claim and verdict judgments an additional 22.7 and 25.0 percentage points toward the Plaintiff, whereas disclosed defendant-only evidence shifted them 15.3 and 15.0 points farther toward the Defendant (all $p < .001$).

Disclosure reduced the influence of defendant-only evidence by 11.8 points for claim judgments ($p < .001$) and 13.2 points for verdict judgments ($p = .001$), but did not significantly reduce the influence of plaintiff-only evidence (both $p > .29$). Crucially, it did not eliminate the influence of either one-sided record. Thus, selectively presented evidence continued to shape legal judgments even when participants were explicitly informed of its one-sided selection.

Study 4: Assigned policy advocacy influences audience judgments despite disclosure

We next moved from legal disputes to contested public policies and asked whether a comparable audience effect arises when an actual advocate constructs the selective record.

Independent reviewers ($N = 288$) each evaluated a dossier—an advocate's selected facts

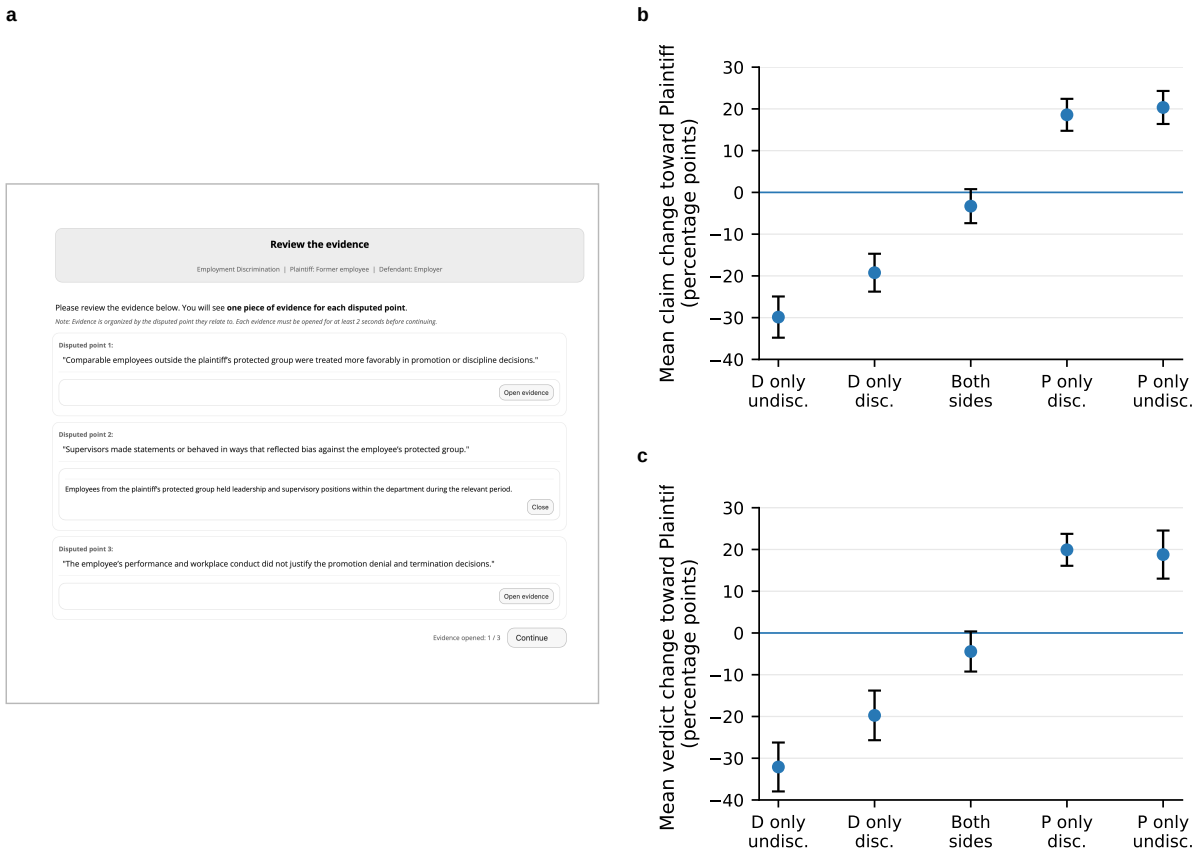


Figure 3: **One-sided legal evidence influences audience judgments despite disclosure.** **a**, Evidence-review page from a one-sided condition, including all three disputed points and one expanded evidence item. Disclosed conditions additionally stated that the displayed evidence came from one side and that opposing evidence could be missing. **b**, Mean disputed-claim updating; even when one-sided selection was disclosed, judgments shifted in favor of the side whose evidence was shown. **c**, Mean verdict updating, which shows the same pattern. P = plaintiff-favoring; D = defendant-favoring; disc. = disclosed; undisc. = undisclosed. Error bars are normal-approximation 95% confidence intervals calculated from participant-level means; positive values indicate movement toward the plaintiff. $N = 399$ participants.

and written argument—produced in Study 6 by an advocate randomly assigned to support or oppose either a minimum-wage increase or carbon pricing. The evidence available to advocates was independently randomized, and their bonus depended on the later reviewer's judgment.

Before seeing the dossier, reviewers rated how warranted they considered five factual claims bearing on the policy, made accuracy-incentivized quantitative estimates of related real-world outcomes, and reported overall policy support. They were then told the advocate's assigned position and evidence-selection task, reviewed the dossier (Figure 4a), and repeated the judgments.

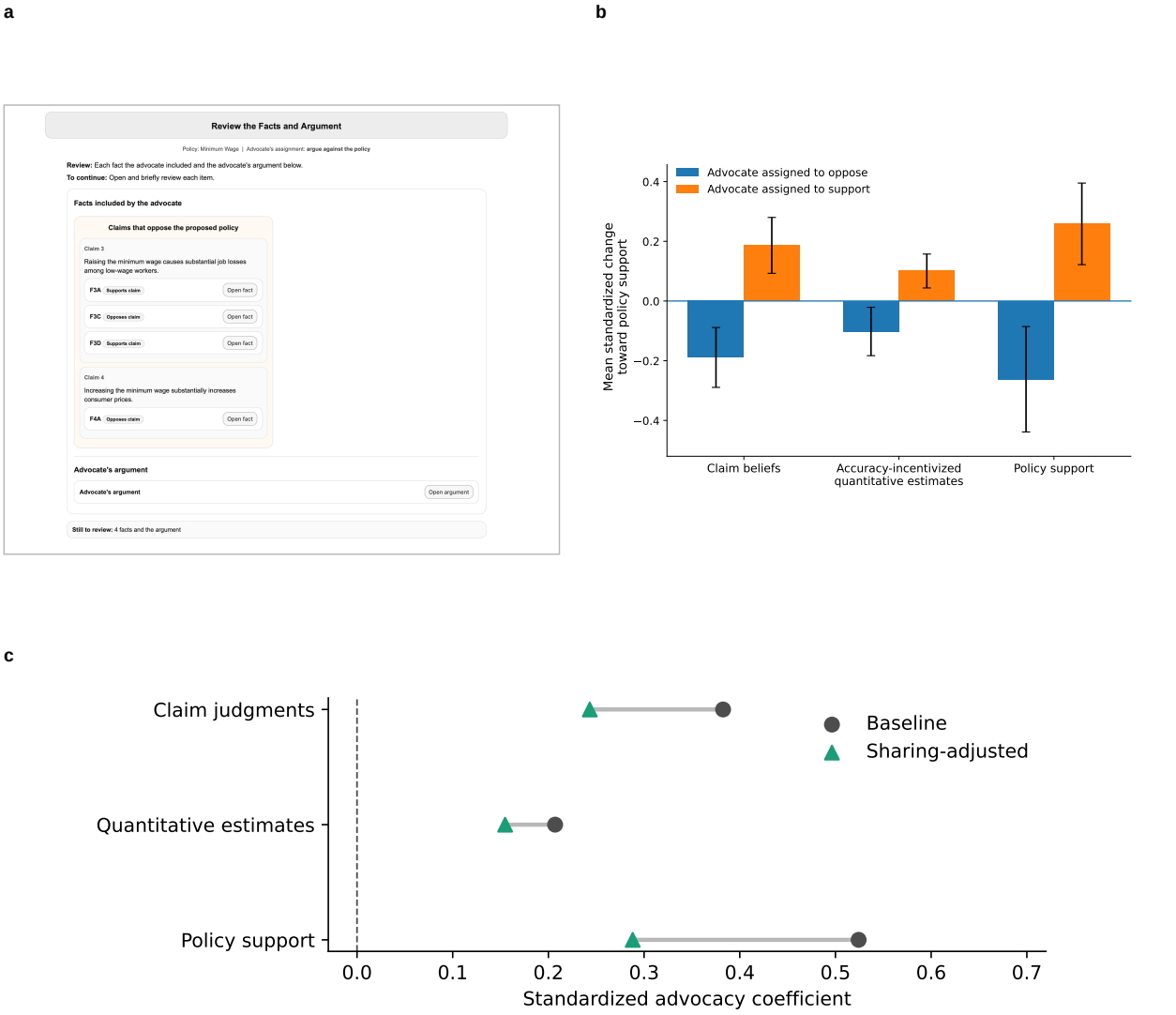


Figure 4: **Assigned policy advocacy influences audience judgments despite disclosure. a**, Reviewer page showing the dossier—the facts selected by the advocate together with the advocate's written argument—and the side the advocate was assigned to support. The argument is available through the collapsed "Open argument" control. **b**, Mean standardized pre-to-post change in claim warrantedness, accuracy-incentivized quantitative estimates, and policy support among reviewers of dossiers produced by advocates randomly assigned to support or oppose the policy. Across all three judgments, reviewers shifted in the direction advocated by the dossier even though they were told that the advocate's position was assigned. Positive values indicate movement toward policy support. Error bars are normal-approximation 95% confidence intervals calculated from reviewer-level means. **c**, Support-minus-oppose advocacy coefficients (standardized units) before and after adjustment for evidence sharing. Connected symbols show the same estimand across the two specifications; movement toward zero after adjustment indicates attenuation. $N = 288$ reviewers.

Despite disclosure, the advocate's assignment shifted all three outcomes (Figure 4b). Controlling for the evidence randomly available to the advocate, support rather than oppose assignment shifted reviewer claim warrantedness by 0.383 SD (95% CI [.246, .519]), quantitative estimates by 0.207 SD (95% CI [.108, .306]), and policy support by 0.524 SD (95% CI [.295, .754]; all $p < .001$). The quantitative effect is particularly informative because reviewers were rewarded for accuracy.

Accounting for selective transmission substantially attenuated these effects (Figure 4c). Controlling for whether and how many pro- and anti-policy facts advocates shared reduced the respective advocacy coefficients by approximately 36%, 25%, and 45% (all $p \leq .004$; Supplementary Information).

Studies 3 and 4 therefore show that selectively constructed records can remain persuasive despite disclosure, whether selection was experimentally imposed or generated by a randomly assigned advocate.

Study 5: Assigned legal advocacy influences advocates' judgments and responses to evidence

We next turned from audiences to advocates themselves, starting with the legal setting.

Participants ($N = 204$) evaluated one of the three legal disputes used in Study 3 and reported baseline beliefs about the case's factual claims and overall verdict before being randomly assigned to advocate for the Plaintiff or Defendant. Independently for each claim, they were shown the plaintiff-favoring fact, defendant-favoring fact, or both. While constructing an argument, participants could open, reference, and attach facts (Figure 5a). We recorded attention through fact openings and viewing time, and evidence use through references and attachments; after submitting the argument, they repeated their judgments and indicated which facts they had considered while making them.

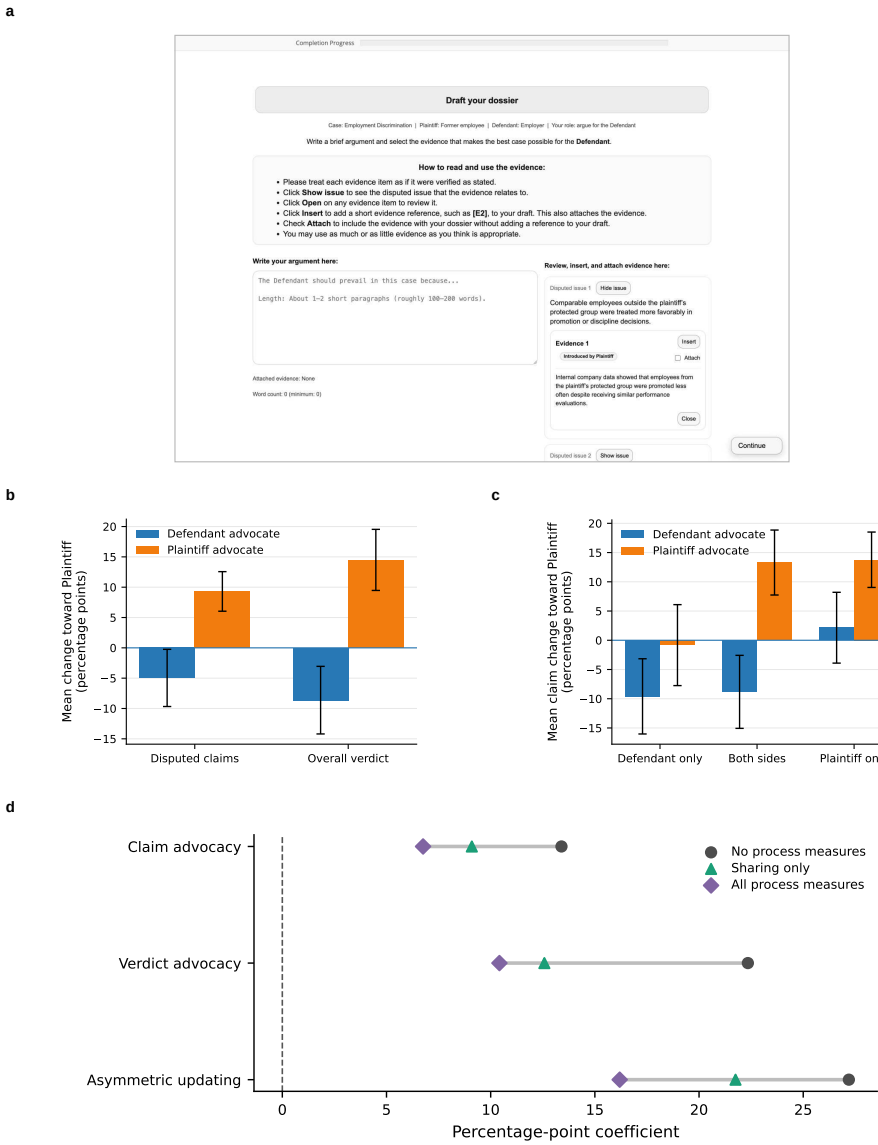


Figure 5: **Assigned legal advocacy influences advocates' judgments and how they respond to evidence. a**, Advocacy-task page with a disputed issue and one evidence item open, showing the assigned side, argument field, and controls for opening, referencing, and attaching evidence. **b**, Mean pre-to-post change in disputed-claim and overall-verdict judgments by randomly assigned advocacy side; both judgments shifted in favor of the assigned side. **c**, Mean claim updating by randomized evidence exposure and advocacy side. Participants responded more strongly to evidence favoring their assigned side, the asymmetric-updating pattern. In **b**–**c**, claim-level observations were averaged within participant and, in **c**, within evidence condition; error bars are normal-approximation 95% confidence intervals across participants; positive values indicate movement toward the plaintiff. **d**, Plaintiff-minus-Defendant coefficients (percentage points) for claim and verdict change and the claim-level asymmetric-updating contrast across three specifications: no process measures, sharing only, and all process measures. "Sharing only" adjusts for the number of facts included with the advocate's argument through an inline reference or attachment. Connected symbols show point estimates for the same estimand across the three specifications; their movement toward zero after adding process measures indicates attenuation. $N = 204$ participants.

Random assignment to an advocacy side shifted beliefs toward that side (Figure 5b). Plaintiff advocates moved 9.3 percentage points toward the Plaintiff across claims, whereas Defendant advocates moved 5.0 points toward the Defendant; the corresponding verdict changes were +14.5 and -8.6 percentage points. Controlling for randomized evidence exposure, Plaintiff rather than Defendant assignment produced differences of 13.4 points in claim updating (95% CI [8.0, 18.8]) and 22.3 points in verdict updating (95% CI [14.9, 29.8]; both $p < .001$).

Advocacy also produced the predicted asymmetric updating in response to randomized evidence (Figure 5c). Plaintiff advocates responded more strongly to plaintiff-favoring evidence, and Defendant advocates more strongly to defendant-favoring evidence. Pooling the two directions yielded a claim-level asymmetric-updating effect of 27.2 percentage points (95% CI [9.1, 45.3], $p = .002$).

The process data showed the same asymmetry. All four measures showed greater processing of advocacy-aligned evidence (Supplementary Information). For example, participants included advocacy-aligned facts with their arguments at rates of 84% rather than 38% and later reported that those facts had come to mind during their final judgments at rates of 83% rather than 39%; advocacy-aligned facts were also opened more often and viewed longer.

Finally, adding all four process measures reduced the claim advocacy effect by 50%, the verdict advocacy effect by 53%, and the claim asymmetric-updating effect by 40% (all $p \leq .001$; Figure 5d). These are associational attenuation estimates, not causal mediation estimates.

Together, these results show that randomized advocacy changed both what participants believed and which evidence most influenced their beliefs. These changes were closely associated with selective processing: participants attended to, shared, and later brought to mind more evidence favoring their assigned side, and accounting for these processing differences substantially attenuated both advocacy-induced belief change and asymmetric updating.

Study 6: Assigned policy advocacy influences advocates' judgments and responses to evidence

Finally, we tested whether these advocacy-induced belief and processing effects generalized to contested public policies.

Participants ($N = 304$) evaluated either a minimum-wage increase or carbon pricing. Before advocacy assignment, they rated how warranted they considered five factual claims bearing on the policy, made accuracy-incentivized quantitative estimates of related real-world outcomes, and reported overall policy support. They were then randomly assigned to argue for or against the policy, with a bonus based on a later reviewer's judgment. Evidence exposure was randomized independently for each claim: participants saw pro-policy evidence, anti-policy evidence, or both before constructing their dossier (Figure 6a).

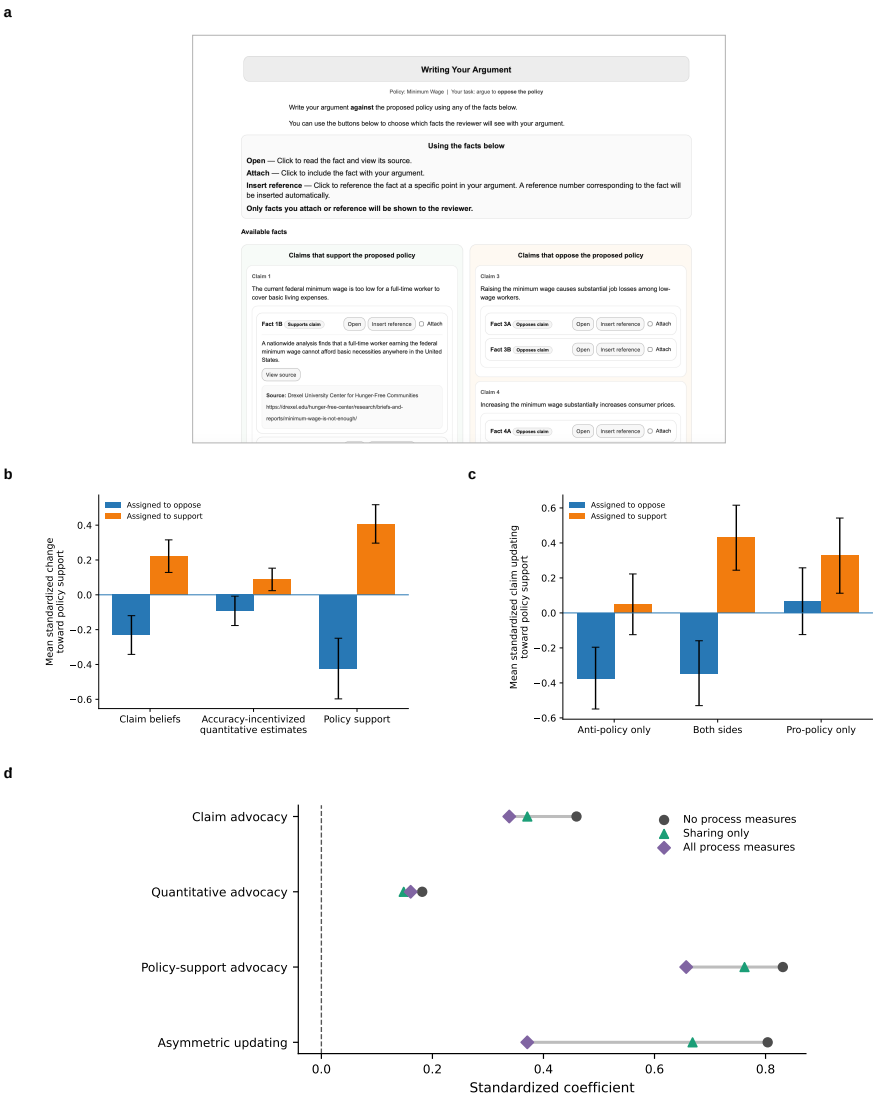


Figure 6: **Assigned policy advocacy influences advocates' judgments and how they respond to evidence. a**, Advocacy-task page showing the available pro- and anti-policy claims, evidence controls, and an opened source. **b**, Mean standardized pre-to-post change in claim warrantedness, accuracy-incentivized quantitative estimates, and policy support by randomly assigned advocacy role; all three judgments shifted in favor of the assigned position. **c**, Mean standardized claim updating for advocacy-aligned claims, shown by randomized evidence condition and advocacy role. Support advocates responded more strongly to pro-policy evidence, whereas oppose advocates responded more strongly to anti-policy evidence, the asymmetric-updating pattern. In **b**–**c**, error bars are normal-approximation 95% confidence intervals calculated from participant-level means; positive values indicate movement toward policy support. **d**, Support-minus-oppose coefficients for the three advocacy effects and the asymmetric-updating estimate across three matched specifications: no process measures, sharing only, and all process measures. The asymmetric-updating estimate combines supporters' greater responsiveness to pro-policy evidence and opponents' greater responsiveness to anti-policy evidence among advocacy-aligned claims. "Sharing only" adjusts for whether any evidence was included with the advocate's argument and the number of facts included. Connected symbols show point estimates for the same estimand across the three specifications; their movement toward zero after adding process measures indicates attenuation. $N = 304$ participants.

Advocacy shifted all three judgments in the assigned direction (Figure 6b). In models controlling for randomized evidence exposure, support rather than oppose assignment shifted claim warrantedness by 0.460 SD (95% CI [.315, .604]), quantitative estimates by 0.182 SD (95% CI [.076, .287]), and policy support by 0.831 SD (95% CI [.625, 1.037]; all $p < .001$). The quantitative effect shows that advocacy shifted judgments even when accuracy was incentivized.

The policy design allowed a more differentiated test of the mechanism underlying our asymmetric-updating prediction. Because our account attributes asymmetric updating to differential processing of case-useful information, it predicts selectivity at both the claim and evidence levels. In the legal study, every claim was an element the Plaintiff had to establish, so Plaintiff advocates benefited from supporting each one and Defendant advocates from undermining each one. In the policy study, advocates instead constructed a brief from claims favoring either side. They therefore had reason to foreground claims favoring their assigned position and leave opposing claims in the background, rather than make those claims salient and then rebut them. We distinguished claim alignment—whether a claim, if true, supported the assigned position—from evidence alignment—whether a piece of evidence directly supported that position. We predicted the strongest asymmetry when both the claim and the evidence were aligned.

That prediction was supported (Figure 6c). Within advocacy-aligned claims, support advocates responded more strongly to pro-policy evidence and oppose advocates more strongly to anti-policy evidence. Pooling these directions yielded an asymmetric-updating effect of 0.804 SD (95% CI [.283, 1.325], $p = .001$). A complementary analysis using all claims likewise showed that this evidence-alignment asymmetry became stronger when the underlying claim was advocacy-aligned (Supplementary Information). The component patterns provided more limited support: pooling across claims, evidence alignment yielded asymmetric updating for claim warrantedness but not the other outcomes; pooling across evidence, claim alignment yielded asymmetric updating only for quantitative estimates (Supplementary Information).

The process data paralleled the belief results. All eight measures showed greater processing of advocacy-aligned evidence and of evidence concerning advocacy-aligned claims (Supplementary Information). For example, participants opened advocacy-aligned evidence bundles 2.15 times on average, compared with 1.68 times for opposing bundles, shared at least one fact from 48.5% versus 21.5% of those bundles, and later reported that at least one fact had come to mind during their final judgments for 69.3% versus 50.9%. They likewise opened evidence about advocacy-aligned claims 3.28 times on average, compared with 1.80 times for other claims, and shared evidence concerning 59.6% versus 26.9% of those claims. Thus, advocates preferentially processed both the claims and the evidence most useful for their assigned case.

Accounting for processing differences attenuated the principal effects (Figure 6d). The full process set reduced the advocacy coefficient by 26% for claim warrantedness ($p < .001$) and 21% for policy support ($p = .010$); the 11% quantitative-estimate attenuation was not significant. The focal asymmetric-updating contrast declined from 0.804 to 0.371, a 54% attenuation ($p = .001$).

We also tested a downstream implication in Study 4. If advocates preferentially process and transmit case-useful information, evidence randomly made available to them should have a larger effect on reviewers when either that evidence or the claim it concerned was useful for the assigned case. These downstream predictions were not generally supported (Supplementary Information), perhaps because advocates compensated through later selection and framing or because reviewers discounted especially one-sided dossiers.

Together, Study 6 extends the Study 5 pattern to contested policy judgments, including accuracy-incentivized quantitative beliefs. Randomized advocacy shifted participants' judgments and produced the strongest asymmetric updating when both the claim and the evidence supported the assigned position. Participants also preferentially processed such case-useful information, and accounting for processing differences substantially attenuated the principal advocacy and asymmetric-updating effects.

Discussion

Across six experiments, we tested a two-step cognitive account of advocacy bias based on selective evidence processing and incomplete adjustment for that selection. Studies 1 and 2 directly demonstrate incomplete correction in tasks where fully selection-adjusted judgments can be calculated. Studies 3 and 4 show a distinct audience-level consequence: disclosure does not neutralize the influence of selectively constructed records, whether the record is imposed experimentally or generated by a randomly assigned advocate. Studies 5 and 6 show that assigned advocacy shifted attention, sharing, and later consideration toward case-useful evidence, and that accounting for these processing differences substantially attenuated the estimated advocacy-bias and asymmetric-updating effects. These findings span stylized and naturalistic settings, legal and policy materials, advocates and audiences, and both persuasion- and accuracy-incentivized tasks.

Several recent papers advance related accounts and provide complementary evidence for parts of this mechanism. Schwardmann, Tripodi, and van der Weele[2] found that randomly assigned debaters shifted toward their assigned positions and disproportionately generated supporting arguments, with selective argument generation partly accounting for the belief shift. Zhang and Rand[40] found comparable belief shifts when participants summarized arguments aligned with an assigned position and when they used those arguments to persuade someone else, suggesting that focused processing of aligned information can shift beliefs even without a motive to persuade. Hunter et al.[41] found that incentives to signal partisan adherence shifted evaluations of novel political evidence and subsequent political beliefs toward a peer's partisan position, with evidence consistent with the belief effect being mediated by evidence evaluation. In a non-advocacy setting, Conlon[42] showed that selective rehearsal changed later recall and beliefs, and that participants did not adequately correct for the way rehearsal had shaped what they remembered.

Others have shown incomplete adjustment to selected evidence in nonstrategic and strategic settings.[33,35,43] For example, in nonstrategic settings, Feiler, Tong, and Larrick[44] studied

an environment in which observed sales were capped by available inventory and therefore concealed the magnitude of demand during stockouts. Although participants knew which observations were censored, their estimates of demand remained too close to the observed sales figures. Enke[45] studied informational homophily: participants disproportionately observed signals from others whose private information was similar to their own. Although they knew that this sorting process made the observed signals unrepresentative, their estimates remained excessively influenced by those signals. In strategic settings, Farina et al.[46] allowed incentivized senders to select which draws from an urn to disclose. Receivers knew the sender's incentives and selection opportunity but treated the selected evidence as more diagnostic than warranted. Degan, Li, and Xie[47] found a similar failure when persuaders chose which of two privately observed signals to reveal. In a legal setting, Koehler and Thompson[48] found that mock jurors responded to the selective nature of evidence from multiple-opportunity searches when the larger sample space was identified, but did not reliably infer the possibility of selection when it was left implicit.

Our findings also accord with the established motivated-reasoning result that directional goals alter information search and evaluation.[5,19,49,50]

Our contribution is to bring these strands together into a single account of advocacy-induced belief change, asymmetric updating, and the influence of selected evidence on informed audiences, while providing further evidence for its component processes and testing the predictions that connect them. Taken together, our findings and this broader evidence suggest that ordinary goal-directed evidence processing, combined with incomplete adjustment for selection, can generate advocacy bias and may also help explain asymmetric updating and the influence of selected evidence on informed audiences.

Methods

Participants, ethics, and preregistration

Participants were recruited online through Prolific. All participants provided informed consent, and all studies were approved by Princeton's Institutional Review Board (Protocol #10857, Computational Cognitive Science). Participants received study-specific base payments and, where described below, could earn additional performance-based bonuses; exact compensation and the paper-wide eligibility criteria are reported in the Supplementary Information.

All six studies were preregistered before collection of the corresponding samples. Preregistration links, target sample sizes, exclusion criteria, and any deviations from the preregistered procedures or analyses are reported in the Supplementary Information. Unless otherwise specified, preregistered directional hypotheses were evaluated using one-sided tests. Detailed analysis specifications are provided in the Supplementary Information, and complete experimental materials and analysis code are available in the public repository.

Study 1: Stick-censoring task

Participants ($N = 50$) completed 27 trials in which they estimated the mean height of 10 sticks. They were told that each height was independently drawn from a uniform distribution over the integers 1–100. The sticks were displayed in ascending order, with between one and nine sticks revealed. In the Tallest-visible condition, the revealed sticks were the tallest in the set; in the Shortest-visible condition, they were the shortest; and in the Random-drop condition, they were randomly selected. Participants knew which selection rule applied and could see the positions of both revealed and hidden sticks. Each participant completed one trial for every combination of selection rule and number of revealed sticks, in random order.

The preregistered analyses tested whether participants adjusted their estimates for the known selection process and whether that adjustment was sufficient relative to a Bayesian benchmark that incorporated the selection rule. Mixed-effects regressions compared estimates across censoring conditions and examined how the effects varied with the number of hidden sticks. We additionally report a non-preregistered proportional-adjustment analysis comparing participants' movement away from the visible-stick mean with the movement required to reach the exact Bayesian benchmark, corresponding to the analysis we subsequently preregistered for Study 2. Full benchmark construction and specifications are reported in the Supplementary Information.

Study 2: Threshold-censoring task

Participants ($N = 116$) completed 30 trials divided into three blocks. On each trial, one of two machines was selected with equal probability. The Low Machine generated values from a normal distribution with mean 230 and the High Machine from an otherwise identical distribution with mean 270 ($SD = 50$). Participants observed only a censored record of the generated values and estimated the probability that the High Machine had generated the record. Estimates were incentivized for accuracy relative to the Bayesian benchmark.

On each trial, the censoring rule was randomly selected. Under Only High, only values above a cutoff were recorded; under Only Low, only values below the cutoff were recorded; and under Random Drop, omission was independent of value. In Block 1, five values were generated and the cutoff was fixed at 250, with both features disclosed to participants. In Blocks 2 and 3, features of the sampling and censoring process were progressively less certain. Before each block, participants could generate example records.

For each trial, we calculated a Naive benchmark that ignored censoring and a Bayesian benchmark that incorporated the task's sampling and censoring process. The preregistered primary analysis measured how far participants adjusted from the Naive toward the Bayesian benchmark, with zero corresponding to no correction and one to complete Bayesian correction. Full benchmark definitions, block-specific procedures, and secondary analyses are reported in the Supplementary Information.

Study 3: Selectively presented legal evidence

Participants ($N = 399$) evaluated one of three anonymized legal disputes concerning product liability, employment discrimination, or consumer fraud. Each case contained several disputed factual claims. For example, one employment-discrimination claim asked whether comparable employees outside the plaintiff's protected group were treated more favorably in promotion or discipline decisions. Each disputed claim had one Plaintiff-favoring and one Defendant-favoring evidence item available.

Participants were randomly assigned to one of five evidence conditions. They saw either evidence from both sides, Plaintiff-favoring evidence only, or Defendant-favoring evidence only; the one-sided conditions additionally varied whether the selective nature of the evidence was explicitly disclosed. Before seeing the evidence, participants reported their beliefs about each disputed claim and the probability that the Plaintiff prevailed. After reviewing the evidence, participants repeated their claim and verdict judgments. In the disclosed conditions, participants were told that they were seeing evidence from only one side. During their post-review judgments, they were also reminded that relevant opposing evidence might be missing and instructed to take into account both the evidence they were shown and the fact that it was one-sided.

Primary outcomes were raw pre-to-post percentage-point changes in claim and verdict judgments, oriented so that positive values indicated movement toward the Plaintiff. Preregistered analyses tested the predicted ordering across the five evidence conditions, whether disclosure reduced the influence of one-sided evidence, and whether one-sided evidence continued to influence judgments after disclosure. Claim-level regressions included disputed-claim fixed effects and participant-clustered standard errors; verdict regressions included case fixed effects and heteroskedasticity-robust standard errors, with one verdict-change observation per participant.

Study 4: Policy-dossier reviewers

Independent reviewers ($N = 288$) each evaluated one dossier produced by an advocate in Study 6. Dossiers concerned either increasing the minimum wage or adopting a carbon-pricing policy and consisted of the advocate's written argument together with the evidence the advocate had chosen to include. Advocates had been randomly assigned to support or oppose the policy.

Before viewing the dossier, reviewers saw a brief summary of the policy debate and five related disputed claims. They then reported how warranted they considered each claim, provided quantitative estimates of measurable outcomes related to those claims, and rated their overall support for the policy. Quantitative estimates were incentivized for accuracy, following related work that elicits quantitative estimates of politically consequential factual beliefs, often with accuracy-contingent payments.[23,51,52]

For example, one minimum-wage item asked: "If the minimum wage were increased by 10%,

approximately what percentage would restaurant prices increase over the following several years?" Participants entered a numerical estimate from 0 to 20%, with responses expressed as the percentage increase in restaurant prices. The accuracy benchmark was 0.36%, based on the reported restaurant-price pass-through associated with a 10% minimum-wage increase in published research. The benchmark and the source and assumptions used to derive it were used for scoring but were not shown to reviewers.

Reviewers were then explicitly informed of the advocate's assigned position and that the advocate had been asked to write a brief argument and select evidence that would be most useful for making the assigned case. A brief comprehension check ensured that reviewers understood the advocate's assignment before proceeding. After reading the argument and selected evidence, reviewers repeated the same judgments. During these follow-up judgments, they were reminded of the advocate's assigned role and, for the quantitative estimates, of the accuracy incentive.

Outcomes were standardized within the corresponding policy-specific claim, quantitative question, or policy. Preregistered analyses estimated the effect of the advocate's randomized assignment while accounting for the evidence the advocate had been randomly shown. Claim-level models included policy-specific claim fixed effects and standard errors clustered by advocate dossier; policy-level models included policy fixed effects. Secondary analyses examined attenuation after accounting for the evidence advocates processed and transmitted.

Study 5: Legal advocacy

Participants ($N = 204$) evaluated one of the three legal disputes used in Study 3. Before receiving an advocacy assignment, they reviewed a brief summary of their assigned case and its disputed claims. They then reported their beliefs about the disputed claims and overall verdict. Participants were subsequently randomly assigned to advocate for either the Plaintiff or the Defendant.

Independently of advocacy assignment, evidence exposure was randomized separately for each claim: participants saw Plaintiff-favoring evidence only, Defendant-favoring evidence only, or both. The evidence they were shown was hidden behind display cards that participants had to click to open, allowing us to record which facts they opened and for how long. While a card was closed, participants could see which claim the fact related to and whether it favored the Plaintiff or Defendant. On the same page on which the evidence was displayed, participants prepared a written argument making the strongest case for their assigned side and could include whichever displayed facts they chose. They could incorporate facts by inserting numbered references inline in their argument or by attaching them without an inline reference. After submitting their argument, participants repeated their claim and verdict judgments and indicated which evidence items had come to mind while making those judgments. We refer to this latter measure as evidence activation.

Preregistered analyses tested whether advocacy assignment shifted participants' beliefs, whether participants responded more strongly to evidence favoring their assigned side,

and whether advocacy changed the processing of advocacy-aligned evidence. Additional preregistered analyses tested attenuation of the advocacy and asymmetric-updating effects after accounting for these process measures. The process measures included the number of fact openings, total fact-viewing time, whether each fact was referenced inline or attached to the dossier, and whether the participant reported considering the fact while making the final judgments. For comparability with Study 3, the main text reports raw percentage-point changes. The preregistered primary analyses used changes in log-odds, claim fixed effects, and participant-clustered standard errors; corresponding verdict analyses included case fixed effects. The preregistered analyses and common-scale robustness analyses are both reported in the Supplementary Information.

Study 6: Policy advocacy

Participants ($N = 304$) evaluated either a proposal to increase the minimum wage or a carbon-pricing proposal. Each policy contained five disputed factual claims. Participants first reported the warrantedness of each claim, quantitative estimates of measurable outcomes associated with the claims, and overall policy support. Quantitative estimates were incentivized for accuracy, following related work that elicits quantitative estimates of politically consequential factual beliefs, often with accuracy-contingent payments.[23,51,52]

Participants were then randomly assigned to advocate either for or against the policy. Their advocacy was incentivized by an additional bonus based on how strongly a later reviewer judged the policy in the assigned direction, similar to outcome-contingent persuasion incentives used in prior experimental work.[51,53,54] Independently for each claim, participants were randomly shown a pro-policy evidence bundle, an anti-policy evidence bundle, or both. Each bundle contained two sourced facts.

For example, for a claim about whether raising the minimum wage reduces employment, one fact stated: "A study of fast-food restaurants in New Jersey and Pennsylvania found little evidence that New Jersey's minimum-wage increase reduced employment." The associated source information identified Card and Krueger and provided a link to the original paper. An opposing fact for the same substantive issue stated that a major review of minimum-wage studies concluded that the evidence generally points to job losses for low-skilled workers after minimum-wage increases. Thus, the evidence items were brief summaries of published findings, with additional source information available on demand.

Participants wrote an argument using any of the available evidence they wished. The interface was similar to Study 5, except that participants could also click to reveal the source associated with each fact, and claims were organized according to whether accepting the claim supported or opposed the proposed policy. The interface recorded the number and duration of openings for each fact and, separately, its source information; whether the evidence was included in the dossier; and whether participants later reported having it in mind while making their final judgments (evidence activation). Participants then repeated the claim, quantitative, and policy-support judgments. These dossiers were subsequently evaluated by the reviewers

in Study 4.

Belief-change outcomes were oriented so that positive values indicated movement toward policy support and standardized within the corresponding policy-specific claim, quantitative question, or policy. Preregistered analyses tested the effect of advocacy assignment, asymmetric responsiveness to randomized evidence, selective processing of advocacy-aligned evidence and claims, and attenuation after accounting for processing differences. Claim-level models included policy-specific claim fixed effects and participant-clustered standard errors; policy-level models included policy fixed effects. Full definitions of the process measures and preregistered specifications are reported in the Supplementary Information.

Data inclusion and statistical analysis

Study-specific completion criteria and preregistered exclusion rules were applied before analysis. Duplicate participation was identified using Prolific identifiers, with later eligible completions excluded. Additional post-preregistration data-quality procedures used in Studies 5 and 6, including the external-assistance/template screen, are described together with complete exclusion accounting in the Supplementary Information.

All primary analyses followed the corresponding preregistrations except where explicitly identified otherwise. Repeated observations were accommodated using participant- or advocate-dossier-clustered standard errors, participant fixed or random effects, or mixed-effects specifications as appropriate to the design. Reported 95% confidence intervals are two-sided normal-approximation intervals. No adjustment for multiple comparisons was applied. We interpret the results as a pattern across prespecified hypotheses rather than treating each nominally significant test as independent confirmatory evidence, and we report all preregistered tests, including nonsignificant and directionally inconsistent results. Complete regression specifications, coefficient tables, robustness analyses, and preregistration deviations are reported in the Supplementary Information.

Software and materials

The experiments were programmed in JavaScript using jsPsych version 8.0 and administered online through cognition.run. Analyses were conducted in Python version 3.13.5 using NumPy, pandas, SciPy, statsmodels, Matplotlib, and PyArrow. Preregistration links are reported in the Supplementary Information.

Data availability

The de-identified data needed to reproduce the reported analyses are available at `https://github.com/Moshe-Hoffman/02_REPRODUCIBILITY_REPOSITORY`.

Code availability

The analysis code, experimental code, and materials needed to reproduce the reported studies are available at `https://github.com/Moshe-Hoffman/02_REPRODUCIBILITY_REPOSITORY`.

Author contributions

All authors contributed to the study designs. M.H. conducted the analyses and wrote the manuscript. E.R. implemented Study 1. M.H. implemented Studies 2–6 with assistance from E.R. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Linda Babcock, George Loewenstein, Samuel Issacharoff, and Colin F. Camerer. Biased judgments of fairness in bargaining. *American Economic Review*, 85:1337–1343, 1995.

[2] Peter Schwardmann, Egon Tripodi, and Joël J. van der Weele. Self-persuasion: Evidence from field experiments at international debating competitions. *American Economic Review*, 112:1118–1146, 2022.

[3] Uri Gneezy, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. Bribing the self. *Games and Economic Behavior*, 120:311–324, 2020.

[4] David E. Melnikoff and Nina Strohminger. The automatic influence of advocacy on lawyers and novices. *Nature Human Behaviour*, 4:1258–1264, 2020.

[5] Charles S. Taber and Milton Lodge. Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50:755–769, 2006.

[6] Daniel Williams. The case for partisan motivated reasoning. *Synthese*, 202:89, 2023. doi: 10.1007/s11229-023-04223-1.

[7] Don A. Moore and Paul J. Healy. The trouble with overconfidence. *Psychological Review*, 115(2):502–517, 2008. doi: 10.1037/0033-295X.115.2.502.

[8] Peter Schwardmann and Joël van der Weele. Deception and self-deception. *Nature Human Behaviour*, 3:1055–1061, 2019.

[9] Miles Hewstone, Mark Rubin, and Hazel Willis. Intergroup bias. *Annual Review of Psychology*, 53:575–604, 2002.

[10] David Pinsof, David O. Sears, and Martie G. Haselton. Strange bedfellows: The alliance theory of political belief systems. *Psychological Inquiry*, 34(3):139–160, 2023. doi: 10.1080/1047840X.2023.2274433.

[11] John T. Jost, Mahzarin R. Banaji, and Brian A. Nosek. A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, 25:881–919, 2004.

[12] Nevin Solak, Maya Tamir, Nebi Sümer, John T. Jost, and Eran Halperin. Expressive suppression as an obstacle to social change: Linking system justification, emotion regulation, and collective action. *Motivation and Emotion*, 45(5):661–682, 2021. doi: 10.1007/s11031-021-09883-5.

[13] George Loewenstein and Andras Molnar. The renaissance of belief-based utility in economics. *Nature Human Behaviour*, 2:166–167, 2018. doi: 10.1038/s41562-018-0301-z.

[14] Jan B. Engelmann, Maël Lebreton, Nahuel A. Salem-Garcia, Peter Schwardmann, and Joël J. van der Weele. Anticipatory anxiety and wishful thinking. *American Economic Review*, 114(4):926–960, 2024. doi: 10.1257/aer.20191068.

[15] Megan K. Smith, Robert Trivers, and William von Hippel. Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 63:93–101, 2017. doi: 10.1016/j.joep.2017.02.012.

[16] William von Hippel and Robert Trivers. The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34:1–16, 2011.

[17] Alice Soldà, C. Ke, Lionel Page, and William von Hippel. Strategically delusional. *Experimental Economics*, 23:604–631, 2020.

[18] Charles G. Lord, Lee Ross, and Mark R. Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109, 1979.

[19] Peter H. Ditto and David F. Lopez. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63:568–584, 1992.

[20] David Eil and Justin M. Rao. The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3:114–138, 2011.

[21] Markus M. Möbius, Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. Managing self-confidence: Theory and experimental evidence. *Management Science*, 68:7793–7817, 2022.

[22] Ben M. Tappin, Leslie van der Leer, and Ryan T. McKay. The heart trumps the head: Desirability bias in political belief revision. *Journal of Experimental Psychology: General*, 146(8):1143–1149, 2017. doi: 10.1037/xge0000298.

[23] Matthew Lilley and Brian Wheaton. Are preconceptions postconceptions? evidence on motivated political reasoning. Working paper, 2024.

[24] Ap Dijksterhuis and Henk Aarts. Goals, attention, and (un)consciousness. *Annual Review of Psychology*, 61:467–490, 2010.

[25] James W. Pichert and Richard C. Anderson. Taking different perspectives on a story. *Journal of Educational Psychology*, 69:309–315, 1977.

[26] Pedro Bordalo, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer. A cognitive theory of reasoning and choice. *Quarterly Journal of Economics*, 141(3):1921–1963, 2026. doi: 10.1093/qje/qjag023.

[27] Fergus I. M. Craik and Endel Tulving. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104:268–294, 1975.

[28] Reid Hastie and Bernadette Park. The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review*, 93:258–268, 1986.

[29] Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–232, 1973.

[30] Nicola Gennaioli and Andrei Shleifer. What comes to mind. *Quarterly Journal of Economics*, 125(4):1399–1433, 2010. doi: 10.1162/qjec.2010.125.4.1399.

[31] Pedro Bordalo, John J. Conlon, Nicola Gennaioli, Spencer Y. Kwon, and Andrei Shleifer. Memory and probability. *Quarterly Journal of Economics*, 138:265–311, 2023.

[32] Timothy D. Wilson and Nancy Brekke. Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, 116:117–142, 1994.

[33] Daylian M. Cain, George Loewenstein, and Don A. Moore. The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *Journal of Legal Studies*, 34:1–25, 2005.

[34] Thomas Graeber. Inattentive inference. *Journal of the European Economic Association*, 21:560–592, 2023.

[35] Kai Barron, Steffen Huck, and Philippe Jehiel. Everyday econometricians: Selection neglect and overoptimism when learning from others. *American Economic Journal: Microeconomics*, 16:162–198, 2024.

[36] Benjamin Enke and Thomas Graeber. Cognitive uncertainty. *Quarterly Journal of Economics*, 138:2021–2067, 2023.

[37] Nicholas Epley and Thomas Gilovich. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, 17(4):311–318, 2006. doi: 10.1111/j.1467-9280.2006.01704.x.

[38] Falk Lieder, Thomas L. Griffiths, Quentin J. M. Huys, and Noah D. Goodman. The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25:322–349, 2018.

[39] Falk Lieder, Thomas L. Griffiths, Quentin J. M. Huys, and Noah D. Goodman. Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, 25(2):775–784, 2018. doi: 10.3758/s13423-017-1288-6.

[40] Y. Zhang and David G. Rand. Self-persuasion does not imply self-deception. *Cognition*, 263:106215, 2025.

[41] Andrew Hunter, Ben Tappin, Charles Efferson, and Ryan McKay. Novel evidence, partisan signalling and divergent political beliefs, 2025. OSF Preprints.

[42] John J. Conlon. Memory rehearsal and belief biases. Preprint, 2026.

[43] Ignacio Esponda and Emanuel Vespa. Endogenous sample selection: A laboratory study. *Quantitative Economics*, 9:183–216, 2018.

[44] Daniel C. Feiler, Jonathan D. Tong, and Richard P. Larrick. Biased judgment in censored environments. *Management Science*, 59:573–591, 2013.

[45] Benjamin Enke. What you see is all there is. *Quarterly Journal of Economics*, 135: 1363–1398, 2020.

[46] Alice Farina, Guillaume R. Fréchette, Alessandro Ispano, Alessandro Lizzeri, and Jacopo Perego. The selective disclosure of evidence: An experiment. *Review of Economic Studies*, 2026. doi: 10.1093/restud/rdag076.

[47] Arianna Degan, Ming Li, and Huan Xie. An experimental investigation of persuasion through selective disclosure of evidence. *Canadian Journal of Economics/Revue canadienne d'économie*, 56:1490–1516, 2023.

[48] Jonathan J. Koehler and William C. Thompson. Mock jurors' reactions to selective presentation of evidence from multiple-opportunity searches. *Law and Human Behavior*, 30(4):455–468, 2006. doi: 10.1007/s10979-006-9021-4.

[49] Kari Edwards and Edward E. Smith. A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, 71(1):5–24, 1996. doi: 10.1037/0022-3514.71.1.5.

[50] William Hart, Dolores Albarracín, Alice H. Eagly, Inge Brechan, Matthew J. Lindberg, and Lisa Merrill. Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4):555–588, 2009. doi: 10.1037/a0015701.

[51] David Hagmann and George Loewenstein. Persuasion with motivated beliefs. Working paper, 2018.

[52] Mike Farjam and Giangiacomo Bravo. Do you really believe that? the effect of economic incentives on the acceptance of real-world data in a polarized context. *Royal Society Open Science*, 11(4):240252, 2024. doi: 10.1098/rsos.240252.

[53] Lucas Coffman and Paul Niehaus. Pathways of persuasion. *Games and Economic Behavior*, 124:239–253, 2020. doi: 10.1016/j.geb.2020.08.008.

[54] Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16:6037, 2025. doi: 10.1038/s41467-025-61345-5.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CognitiveAccountofAdvocacyBiasSINHB.pdf